A Density Function for Cluster Analysis

Susan Laflin.

Computer Centre
University of Birmingham

Abstract

The S.A.H.N. methods in the Clustan package are applied to the objects in a similarity matrix whether any obvious clusters are present in the data or not and the interpretation of the resulting dendogram can be extremely difficult.
In this paper the author suggests an approach to decide whether the data in a given similarity matrix is suitable for cluster analysis or not. Much work remains to be done but the method is aimed at deciding whether significant clusters are present in the data and only continuing with the clustering if the form of the data justifies this.

Introduction.

The motivation for this study arose from a consideration of the SAHN methods within the Clustan package. This works on a matrix of similarities, either supplied as data or calculated within the package and processes the data starting with clusters each containing one object and merging clusters together at suitable similarity levels until it concludes with one cluster containing all the objects. This process is carried out whether or not the data contains any recognisable clusters at all and produces vast amounts of output of which the dendogram is probably the most useful. Running Clustan uses a comparatively large amount of computer time and storage and interpreting the results requires a lot of time and effort from the archaeologist. Accordingly I attempted to prevent this waste of resources by devising a simple way in which the data could be checked in advance to see whether it contains clusters and only processed by the Clustan package if they were shown to be present.

I restricted my ideas to the case where we have a similarity or distance matrix, since the original data matrix is not always available. If we have the original data, then the choice is even wider: in addition to calculating a suitable distance matrix we may also use methods such as principal component analysis to examine the data.

It is also possible, without loss of generality, to consider a distance matrix, scaled to lie in the range 0.0 to 1.0. Any other distance matrix can be multiplied throughout by a suitable constant to place it in this range. Any similarity matrix can be similarly scaled and then converted by setting $d_{ij} = 1.0 - s_{ij}$.

Identification of Clusters.

In order to test for the existence of clusters, we have to decide what we mean by the term. We all have a vague idea of a group of objects, all very similar to each other and distinct from the other objects in the study. To obtain a definition which can be used to test the objects within a distance matrix, let us consider the two dimensional case. Figure 1 shows a number of objects plotted against the two attributes used to

describe them. Case 1(a) shows clear examples of clusters, some being nice compact circular ones while others are more elongated. Each cluster is typified by an area of "high density" (in which there are a large number of objects per unit volume of attribute space) surrounded by a "moat" of "low density" (with few or no objects) and additional objects beyond this. It is this high density centre, surrounded by a low density moat, surrounded in turn by other regions of medium density for which we wish to test.

Case 1(b) shows a completely regular distribution with the same number of items per unit area throughout. This is in its own way as artificial as the clusters and would usually be interpreted as evidence of human intervention.

Case 1(c) is a random distribution, which may in places show some hints of clustering but such clusters are not significant and we wish to exclude them. This is generally assumed to show the natural state without the influence of human organisation. For each object i, we can study the distance matrix and obtain a count of $N_i(0.1)$, the number of other objects within a small distance, say 0.1 of this one. We are interested in those objects i with large values of N_i since these will lie near the centre of a cluster if any objects do. Let us take the object with the maximum value of $N_i(0.1)$. We now know the number of objects that lie within a distance 0.1 of object i but have no idea of the direction in which they lie, so in the two-dimensional case we may say they lie somewhere within a circle of radius 0.1. This gives a value of density:

$$D_i(0.1) = N_i(0.1) / \pi\ 0.1^2$$

Similarly for the same object if $N_i(0.2)$ is the number of objects with distance between 0.1 and 0.2,

$$D_i(0.2) = N_i(0.2) / \pi(0.2^2 - 0.1^2)$$

and so for each value up to 1.0.
Then for the object i, we may plot this density function $D_i(r)$ against r. If the object lies at the centre of a compact circular cluster, we will get high values of D_i for the cluster, low or zero values for the moat and higher values again outside it.

If it is a multi-dimensional case of dimension k we are dealing with concentric hyperspheres and so the expression should become

$$D_i(r) = N_i(r) / \pi(r^k - (r - 0.1)^k)$$

and the same behaviour should become apparent.

Future Work.

A number of questions still remain unanswered. For example, is the density function really needed or would the histogram of $N_i(r)$ give the required information? How big a step do we need between the cluster and the moat to distinguish it from the random case? How many dimensions do we need to consider? Is 0.1 the correct step size in searching for clusters? My next task is to carry out computations to investigate these questions and I hope to report my results next year.

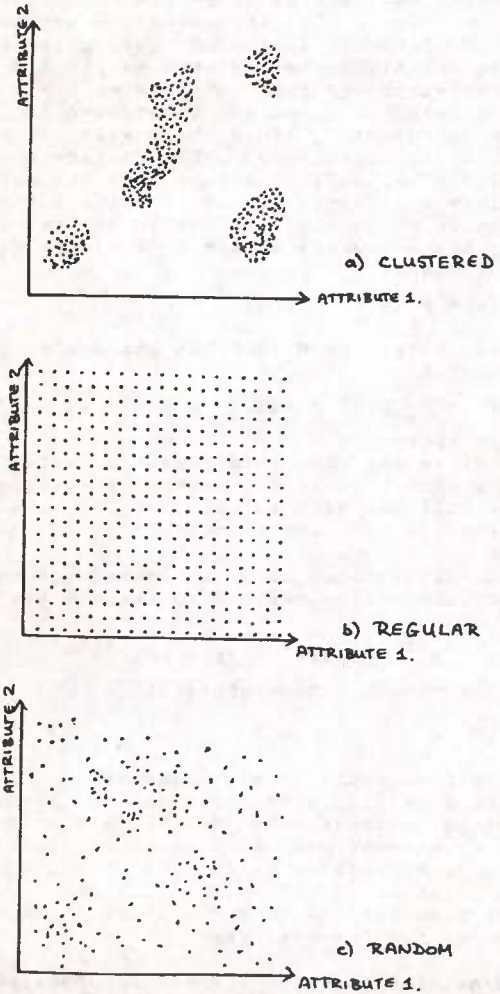
I have spent some time checking for similar published work. The technique of "mode analysis" in the Clustan package uses a

related idea, but has no clear recommendations on how to distinguish clusters from random distributions. The textbook by Everitt mentions some similar work which I have not yet been able to study in detail.

References:

Users Guide to Clustan 1A. University of Bradford.
Cluster Analysis. Brian Everitt, 1974 Heinemann.

FIGURE 1.



THE PROCESSING OF THE POTTERY FROM THE WROXETER BATHS BASILICAN AREA EXCAVATIONS OF 1966-1977

Pamela V Clarke Inspectorate of Ancient Monuments

The pottery is derived from the latest archaeological layer overlying features from the latest occupation of the site, and is being processed with the aim of producing a detailed numerical analysis, showing the composition and distribution of all of the pottery groups.

In setting out the reasons for the study taking this form it it first necessary to give a brief archaeological summary. The following has been extracted from the site interim reports chiefly Barker (1973).

A great deal of the 6,800 sq m site has been destroyed by activity on the site subsequent to the abandonment of it, some in antiquity and some by previous excavators, however enough remains to show that there were many periods of occupation on the site of the basilican building. Dame Kathleen Kenyon (1940) suggests the date 350 AD as the probable date of abandonment and demolition of the basilica, and it is the subsequent occupation of its site, with which we are concerned. At the eastern end of the site the emptying of old excavation trenches has shown that there are 15 discernible layers between the floor of the basilica and the latest building levels (although in other areas there are not as many as this).

In this latest period there is evidence that the basilica area was completely redeveloped, with a planned complex of timber framed buildings, some of them very large, and with a symmetrical plan suggestive of classical prototypes.

Such a drastic re-organisation of the city centre needed wealth, strong motivation and a high degree of organisation, it has all the hall-marks of Roman public works. There are no signs of violent destruction of this occupation, nor any signs of hurried abandonment, the major buildings appear to have been dismantled.

The street running east-west and bounding the north side of the insula had had the surface stripped of metalling, a procedure which involved the removal of hundreds of tons of consolidated pebbles, the lowered surface of the road was then covered with a very fine gravel, produced by sifting the rubble from the demolished basilica. It has been suggested that the former street then became a covered arcade.

A bow sided building has been recorded at the extreme eastern end of the site, the first to be found in association with a Roman site, though the actual date of its construction is not clear.

As well as the stratigraphic evidence of a long period of post 350 AD occupation, other indications include the discovery of a coin dated to AD 367, found stratified beneath building 6, and a tombstone, dedicated to an Irish king, Cunorix, dated by Professor Keneth Jackson to circa AD 460–480, which may indicate that mercenaries were present.

Thus it can be appreciated that the pottery under consideration, derived from a layer overlying the whole site, and directly underlying the plough soil comprises a potentially important group of material, containing material derived from the post 350 AD occupation deposits as well as considerable quantities of residual pottery.

While an important group however, there are a number of practical problems associated with the processing and publication of the material. The current estimate of the probable quantity of pottery to be processed is over 10,000 sherds. This material has accumulated over 400 years of occupation of the site, and the late material, about which we know least, probably forms only a very small proportion of the total.

Bearing in mind the problems, it was decided that the only satisfactory way of examining this material was by drawing up a type series of forms and fabrics present, and quantifying the pottery according to this.

The site recording is based on a 10 m grid, sub divided into 16ths, and therefore most material can be located to within 2.5 m of its find position, but as the material can be regarded as a single group the pottery is not retained in its grid grouping once washing marking and bagging is complete.

When further processing is begun the material is sorted according to a general classification into white ware, red ware, grey ware, black burnished ware, colour coat, calcite grit, samian and any other general categories that occur eg amphora. Mortaria are not separated out but sorted into their appropriate general group.

Each of these is then examined in turn to identify individual fabric groupings within the general group. How this is done varies with the groups, some can be easily divided by simple examination of the sherds, others need sorting into vessel forms first, often it is a simple question of removing one particularly obvious fabric group, and examining the remaining material for further sub division. This is the most difficult part of the work, and the most time consuming, it is based entirely on visual examination of the sherds, usually with the help of a hand lens. None of the divisions have yet been tested scientifically, but geological differences are not the only ones which need consideration and visual differences should always be taken note of initially, even if they later prove to be of no significance.

When the pottery has been classified according to fabric and form it is quantified by counting and weighing each sherd or group of sherds having a common site location. (See Appendix).

No attempt is made to estimate the number of vessels present, as it is felt that the material must have been moved over the site many times during the 400 years or so, in which it accumulated and so the final location of all the sherds from one individual vessel is simply not relevant. Joins are noted where they are found to exist, but it is not considered cost effective to systematically check for joins or sherds of the same vessel.

We are fortunate because of the quantity of material involved that we have access to computer facilities, and I am grateful to Sue Laflin, for the work that she is doing in this connection. A simple number and letter code is used to record the information on each sherd onto a punch card, and computer files are built up as the classification of the pottery proceeds, in the following format

1. fabric type and sub-group within the type;
2. vessel form, this can be general or specific, depending what part of a vessel the sherd comes from;
3. any other features of the pot, eg surface treatment, decoration, repair etc;
4. grid square;
5. layer and/or feature;
6. number of sherds;
7. weight of sherds in grams.

This information is recorded using a simple letter and number code, and recorded on a computer form for punching onto cards.

How simple the code is can be illustrated by explaining the following computer coded sherd description:

1	2	3	4	5	6	7
BBB	DB	E	A2	1-2	1	25

First the fabric group, this is designated by the letters BBB, the first 2 letters indicate that the general group is Black Burnished ware, the third letter also a B indicates that it is of the second fabric group recognised, the letters DB indicate that it is a dish base, it not being possible to identify the form further, the letter E, which comes next indicates that there is decoration on the external surface of the base, (a U would indicate that the upper or internal surface was decorated) A2 is the grid square, actually 2 alpha, a grid square external to

grid square 2, but Greek characters are not available on the punch card. The sherd comes from layer 1-2, 1 indicates that one sherd is involved and 25 indicates that the sherd weight is 25 grams.

When a batch of cards has been punched it is fed into the computer, and forms the basis of a series of files, one for each general category of material. It will eventually be possible to produce lists of material in various formats for preparation of the excavation and pottery reports, facilitating the complete publication of a very large group of material.

The ultimate aim is to devise a system which will produce a detailed analysis of the pottery, and make it possible to distinguish with some degree of confidence, and with a measure of objectivity, the groups of material most frequently associated with the late occupation of the site. An attempt will also be made to interpret the evidence embodied in the on site distributions of all of the groups of material.

In order to do this the following information will first be produced

1. the quantity of material present in each fabric group;
2. proportions present in relation to each other;
3. number of different vessel forms present in the group;
4. distribution of vessel forms through the fabric groups;
5. relative proportions of types occurring;
6. distribution of all the pottery types over the site;
7. distribution of each pottery fabric and form over the site;
8. correlations and divergences in distribution of fabric and form types;
9. differences and similarities in distribution of forms where the fabric is the same.

It is intended that the on site distribution of the pottery will be correlated with the distribution of archaeological features when phasing has been carried out. This should make it possible to identify wares occurring only in association with late features on the site. Further work on pottery from earlier features and layers will make it possible to identify the point at which particular groups of material first appears on the site, and it will therefore eventually be possible to make use of material so identified as a useful indicator of date on other areas of the town site.

The excavation is taking place on a monument in Guardianship, not as a rescue excavation, and as it is primarily a research project, a more innovative approach to pottery processing than is usually possible has been adopted, in order to test methods and theories with application to the particular problems posed by this site, but also of possible use elsewhere. The main aim has been to attempt to

devise a system which made it possible to distinguish with confidence, and with some degree of objectivity, the residual material, from that associated with the latest occupation of the site. Other unsuspected advantages of the method are emerging all the time as work progresses.

It is possible to produce the results of batches of data from groups of work which have been completed, including histograms showing the composition of fabric groups currently classified, and have lists showing the ditribution of the processed material over the site.

This has reassuringly shown that so far the different wares do have differential distributions related to the archaeology of the site.

It also appears that fragment size does have some correlation with on-site location. The mass of material and the size of the area offer the opportunity to explore in more detail than is usually possible the relationship of the material to the area in which it was found.

On a practical note processing does not necessarily take any longer than would an examination of the material using a less detailed recording system, and it is increasingly being required of pottery reporters that some form of quantification of materials is provided. This system could profitably be adopted elsewhere. The method can be applied to any group of pottery, large or small and access to a computer is not absolutely necessary, a card index, punch card system or information record sheets could be used, for smaller quantities of material.

As work proceeds we are hopeful that the sort of information we require, and much more, will emerge from this detailed pottery study.

REFERENCES

Barker P A 1973 — EXCAVATIONS ON THE SITE OF THE BATHS BASILICA AT WROXETER 1966-73. University of Birmingham

Kenyon Dame Kathleen — 'Excavations at Viroconium 1936-37' ARCHAEOLOGIA vol LXXXVIII (1940) pp 175-227

APPENDIX

QUANTIFICATION OF POTTERY

Pottery workers are increasingly asked for quantification of the material on which they are working. In view of the continuing debate on the validity and usefulness of the various methods of doing this it was decided to test the respective merits of weight and number by both weighing and counting sherds.

What has emerged from doing this is that the recording of both of these characteristics can provide a sensitive indication of archaeological differences on site. In the group currently under consideration large sherds tend to congregate in particular areas, while small fragmentary sherds also occur together and in different areas to the large fragments. This seems to be a reflection of differences of use and disturbance of areas of the site, rather than of the degree of fragility of the various classes of pottery represented. I would therefore suggest that where, as in this case, examination of the material as sherds proves the only possible method of quantification, it will prove useful to record both weight and number, as the degree of fragmentation of sherds from any one group, compared with that of other groups from the site may be of archaeological significance.